

This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

**As rescanning documents *will not* correct images,
please do not report the images to the
Image Problem Mailbox.**



PATENT ABSTRACTS OF JAPAN

(11) Publication number: **10260979 A**(43) Date of publication of application: **29.09.98**

(51) Int. Cl.

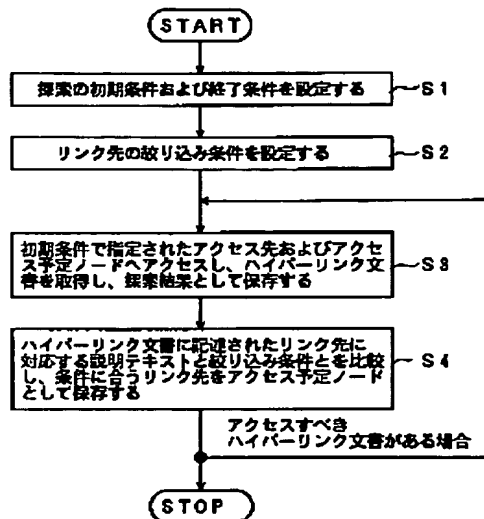
G06F 17/30(21) Application number: **09065106**(22) Date of filing: **18.03.97**(71) Applicant: **NIPPON TELEGR & TELEPH
CORP <NTT>**(72) Inventor: **MATSUO HIROSHI**(54) **INFORMATION COLLECTING METHOD AND
DEVICE**

COPYRIGHT: (C)1998,JPO

(57) Abstract:

PROBLEM TO BE SOLVED: To efficiently collect information by comparing an explanation text that corresponds to a link destination described on a hyperlink document with a narrowing condition and holding the link destination that coincides with the narrowing condition as an access prearranged node.

SOLUTION: A start condition of search and an end condition of search are set (step S1), a narrowing condition of a link destination is set (step S2), a hyperlink document is acquired by accessing an access destination that is designated with the start condition and an access prearranged node, and the acquired hyperlink document is stored as a search result (step S3). An explanation text that corresponds to a link destination which is described on the hyperlink document is compared with a narrowing condition, and a link destination that coincides with the narrowing condition is held as an access prearranged node (step S4). The above processing is performed of entire hyperlink document to be accessed.



(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平10-260979

(43) 公開日 平成10年(1998) 9月29日

(51) Int.Cl.⁶

G 0 6 F 17/30

識別記号

F I

G 0 6 F 15/40

3 7 0 A

3 1 0 F

15/403

3 2 0 D

3 5 0 C

15/419

3 2 0

審査請求 未請求 請求項の数4 O L (全 9 頁)

(21) 出願番号

特願平9-65106

(22) 出願日

平成 9 年 (1997) 3 月 18 日

(71) 出願人 000004226

日本電信電話株式会社

東京都新宿区西新宿三丁目19番2号

(72) 発明者 松尾 比呂志

東京都新宿区西新宿三丁目19番2号 日本

電信電話株式会社内

(74) 代理人 弁理士 伊東 忠彦

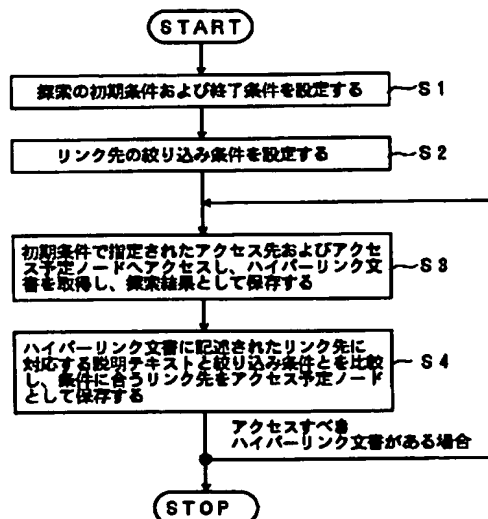
(54) 【発明の名称】 情報収集方法及び装置

(57) 【要約】

【課題】 すべてのリンク先をアクセスするのではなく、絞り込み条件を満たすリンク先のみをアクセスすることによって、ユーザの所望する可能性が高い文書に対するアクセスに限定して、効率よく情報を収集することが可能な情報収集方法及び装置を提供する。

【解決手段】 ハイパーリンク文書のリンク先を辿る時に、リンク情報に対応した説明テキストを絞り込み条件によって判定し、リンク先を絞り込むことによって、有効なハイパーリンク文書を効率的に収集する。

本発明の原理を説明するための図



【特許請求の範囲】

【請求項1】 ネットワーク上に分散されたサーバ上にハイパーリンク文書が分散蓄積されている場合の情報収集を支援するための情報収集方法において、探索の初期条件及び終了条件を設定し、リンク先の絞り込み条件を設定し、前記初期条件で指定されたアクセス先及びアクセス予定ノードへアクセスし、前記ハイパーリンク文書を取得し、当該ハイパーリンク文書を探索結果として保持しておき、前記ハイパーリンク文書に記述されたリンク先に対応する説明テキストと前記絞り込み条件とを比較し、該絞り込み条件に合致するリンク先をアクセス予定ノードとして保持する処理を、アクセスすべき全ハイパーリンク文書について行うことを特徴とする情報収集方法。

【請求項2】 前記リンク先の絞り込み条件として指定されたキーワードを、語の概念の関係を表すシソーラス辞書に基づいて展開し、照合対象語として設定し、前記ハイパーリンク文書に記述されたリンク先に対応する説明テキスト中に、前記照合対象語が含まれるリンク先をアクセス予定ノードとする請求項1記載の情報収集方法。

【請求項3】 ハイパーリンク文書のリンク先を辿って文書の自動収集を行う情報収集装置であって、探索の初期条件及び終了条件を設定する探索条件設定手段と、リンク先の絞り込み条件を設定する絞り込み条件設定手段と、ハイパーリンク文書に記述されたリンク先に対応する説明テキストを前記絞り込み条件に基づいて判定し、条件に合うリンク先をアクセス予定ノードとするアクセス候補絞り込み手段と、前記初期条件で指定されたアクセス先及び前記アクセス予定ノードへアクセスしてハイパーリンク文書を取得し、該ハイパーリンク文書を探索結果として保存するハイパーリンク文書アクセス手段と、前記初期条件及び前記終了条件に基づいて、前記ハイパーリンク文書アクセス手段及び前記アクセス候補絞り込み手段を繰り返し起動する探索管理手段とを有することを特徴とする情報収集装置。

【請求項4】 語の概念の関係を表すシソーラス辞書を更に有し、前記アクセス候補絞り込み手段は、前記絞り込み条件として指定されたキーワードを、前記シソーラス辞書に基づいて展開して照合対象語として設定する絞り込み条件設定手段と、前記ハイパーリンク文書の記述されたリンク先に対応する説明テキスト中に前記照合対象語が含まれるリンク先をアクセス予定ノードとする絞り込み手段を含む請求項3記載の情報収集装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、情報収集方法及び装置に係り、特に、ハイパーリンク文書のリンク先を辿って文書の自動収集を行う情報収集方法及び装置に関する。詳しくは、ネットワーク上に分散されたサーバ上にハイパーリンク文書が分散蓄積されている場合の情報収集を支援するための情報収集方法及び装置に関する。

【0002】

10 【従来の技術】 従来の情報収集装置について説明する。図8は、従来の情報収集装置の構成を示す。従来の情報収集装置は、探索条件設定部C1、探索部C2及び探索結果表示部C3から構成される。探索部C2は、ハイパーリンク文書アクセス部C21、アクセス予定ノード抽出部C22、終了条件判定部C23より構成される。

【0003】 図9は、従来の情報収集装置の動作を示す。ステップ10) まず、探索条件設定部C1により、初期アクセスノードと最大ノード深さを探索条件として設定する。ステップ11) 初期条件として、初期アクセスノードをカレントノードとする。

20 【0004】 ステップ12) ハイパーリンク文書アクセス部C21により、カレントノードにアクセスし、ハイパーリンク文書を取得し、探索結果として保存する。ステップ13) 終了条件判定部C23により、カレントノードが最大深さにあるかを調べ、最大深さにある場合には、次のステップ15に移行する。最大深さにない場合には、ステップ14に移行する。

【0005】 ステップ14) アクセス予定ノード抽出部C22により、ハイパーリンク文書に記述されたリンク先をアクセス予定ノードとして保存する。ステップ15) 終了条件判定部C23により、未処理のアクセス予定ノードがあるかを調べ、あれば、それをカレントノードとする。未処理のアクセス予定ノードがある場合には、ステップ12に移行する。

【0006】 ステップ16) 探索結果表示部C3により探索結果を表示する。以上の処理により、初期アクセスノードから最大ノード深さまでのリンクされたハイパーリンク文書を全て取得することができる。

【0007】

40 【発明が解決しようとする課題】 しかしながら、上記従来の装置では、リンクされたすべてのハイパーリンク文書にアクセスするため、指定する最大ノード深さが深い場合には、大量の文書を取得してしまう。1文書あたりn個のリンク先があり、最大ノード深さをmと指定した場合、nのm乗個となり、探索の深さを大きくすると、膨大な量の文書をアクセスすることになる。例えば、n=5、m=5とすると、5階層目で3125個の文書をアクセスすることになる。

50 【0008】 しかしながら、ユーザが欲しい情報は、これら文書のうちごく一部であることが多い。このため、

検索処理を施して収集した文書を絞り込んだり、人手で必要な文書を捜し出すなどの作業に多くの労力を要する。また、ネットワーク上に分散配置された文書を大量にアクセスすると、ネットワークへの大きな負担がかかるという問題がある。

【0009】本発明は、上記の点に鑑みなされたもので、すべてのリンク先をアクセスするのではなく、絞り込み条件を満たすリンク先のみをアクセスすることによって、ユーザの所望する可能性が高い文書に対するアクセスに限定して、効率よく情報を収集することが可能な情報収集方法及び装置を提供することを目的とする。

【0010】

【課題を解決するための手段】図1は、本発明の原理を説明するための図である。本発明は、ネットワーク上に分散されたサーバ上にハイパーリンク文書が分散蓄積されている場合の情報収集を支援するための情報収集方法において、探索の初期条件及び終了条件を設定し（ステップ1）、リンク先の絞り込み条件を設定し（ステップ2）、初期条件で指定されたアクセス先及びアクセス予定ノードへアクセスし、ハイパーリンク文書を取得し、取得したハイパーリンク文書を探索結果としての保存を行い（ステップ3）、ハイパーリンク文書に記述されたリンク先に対応する説明テキストと絞り込み条件とを比較し、該絞り込み条件に合致するリンク先をアクセス予定ノードとして保持する（ステップ4）処理を、アクセスすべき全ハイパーリンク文書について行う。

【0011】また、本発明は、リンク先の絞り込み条件として指定されたキーワードを、語の概念の関係を表すシソーラス辞書に基づいて展開し、照合対象語として設定し、ハイパーリンク文書に記述されたリンク先に対応する説明テキスト中に、照合対象語が含まれるリンク先をアクセス予定ノードとする。

【0012】図2は、本発明の原理構成図である。本発明は、ハイパーリンク文書のリンク先を辿って文書の自動収集を行う情報収集装置であって、探索の初期条件及び終了条件を設定する探索条件設定手段1と、リンク先の絞り込み条件を設定する絞り込み条件設定手段2と、ハイパーリンク文書に記述されたリンク先に対応する説明テキストを絞り込み条件に基づいて判定し、絞り込み条件に合うリンク先をアクセス予定ノードとするアクセス候補絞り込み手段3と、初期条件で指定されたアクセス先及びアクセス予定ノードへアクセスしてハイパーリンク文書を取得し、該ハイパーリンク文書を探索結果として保存するハイパーリンク文書アクセス手段4と、初期条件及び終了条件に基づいて、ハイパーリンク文書アクセス手段4及びアクセス候補絞り込み手段3を繰り返し起動する探索管理手段5とを有する。

【0013】また、本発明において、語の概念の関係を表すシソーラス辞書を更に有し、アクセス候補絞り込み手段3は、絞り込み条件として指定されたキーワード

を、シソーラス辞書に基づいて展開して照合対象語として設定する絞り込み条件設定手段と、ハイパーリンク文書の記述されたリンク先に対応する説明テキスト中に照合対象語が含まれるリンク先をアクセス予定ノードとする絞り込み手段を含む。

【0014】このように、本発明による情報収集方法及び装置では、アクセスしたハイパーリンク文書に記述されたリンク先に対応する説明テキストと、絞り込み条件設定手段により設定されている絞り込み条件とを比較することにより、合致するリンク先をアクセス予定ノードとして絞り込むことにより、不要な文書へのアクセスを回避して、ユーザが所望する情報が含まれる可能性がある文書を効率よく収集することが可能となる。

【0015】

【発明の実施の形態】図3は、本発明の情報収集装置の構成を示す。同図における情報収集装置は、探索条件設定部1、絞り込み条件設定部2、アクセス候補絞り込み部3、ハイパーリンク文書アクセス部4、探索管理部5、アクセス管理テーブル6、取得文書蓄積部7、処理要求受付部8、探索結果表示部9、照合対象語蓄積部10、及びシソーラス辞書11から構成され、当該情報収集装置は、ネットワーク12を介してサーバ群13に接続される。

【0016】アクセス管理テーブル6は、アクセス予定の文書の識別子（以下、アクセス予定ノードと呼ぶ）等を保存して、文書のアクセスの管理に使用するテーブルである。当該アクセス管理テーブル6には、アクセス予定ノード、処理結果、ノードの深さ等の項目が定義される。アクセス予定ノードには、アクセス予定として決定された文書の識別子が記述され、処理結果には、当該文書が取得されたか否かによって、『処理済』か『未処理』が記述され、ノード深さには、開始文書を0として、開始文書から何個のリンクで結ばれているかを示す個数が記述される。

【0017】取得文書蓄積部7は、ハイパーリンク文書アクセス部4によって取得された文書を蓄積しておく。処理要求受付部8は、ユーザとの入出力インタフェースであり、情報入出力の機能を持つか、ネットワークを介して端末と接続され、ユーザからの探索条件設定要求、絞り込み条件設定要求、探索要求、探索結果出力要求を受付、各々、探索条件設定部1、絞り込み条件設定部2、探索管理部5、探索結果表示部9へ仲介する。

【0018】探索結果表示部9は、処理要求受付部8からの要求に応じて、取得文書蓄積部7に蓄積された文書を出力する。照合対象語蓄積部10は、アクセス候補絞り込み部3で参照できるように、絞り込み条件設定部2によって設定される照合対象語を保存しておく。シソーラス辞書11は、語が表す概念の関係を示す情報を蓄積した辞書であり、絞り込みを行うための知識として使用される。

【0019】探索条件設定部1は、処理要求受付部8を介して探索条件の要求を受け取り、探索条件の初期条件として開始文書を設定し、終了条件として、最大ノード深さを設定する。絞り込み条件設定部2は、処理要求受付部8を介して絞り込み条件設定要求を受け取り、絞り込み条件を設定する。絞り込み条件の指定の方法として、キーワードを指定する方法、条件式を指定する方法、照合パターンを指定する方法等、各種指定方法を適用することができる。

【0020】探索管理部5は、処理要求受付部8を介して探索要求を受け取ると、アクセス管理テーブル6を参照・更新しながら、探索条件で指定された終了条件に達するまで、アクセス候補絞り込み部3と、ハイパーリンク文書アクセス部4とを繰り返し起動して探索を行う。アクセス候補絞り込み部3は、ハイパーリンク文書中から、リンク先に対応する説明テキストを抽出し、照合対象語蓄積部10に保存された照合対象語が説明テキスト中に含まれるリンク先をアクセス予定ノードとして抽出し、探索管理部5を介して、アクセス管理テーブル6へ登録する。

【0021】ハイパーリンク文書アクセス部4は、探索管理部5からアクセスすべきノードの情報を受け取り、ネットワーク12を介して、サーバ群13をアクセスして、該当するハイパーリンク文書を取得し、取得文書蓄積部7へ保存する。図4は、本発明の情報収集方法の一連の動作を示すフローチャートである。

ステップ101) 探索条件設定部1は、探索条件の初期条件として、開始文書が設定され、終了条件として最大ノード深さが設定される。

【0022】ステップ102) アクセス候補絞り込み部3は、開始文書をアクセス予定ノードとして、アクセス管理テーブル6へ登録する。

ステップ103) 絞り込み条件設定部2は、処理要求受付部8からの絞り込み条件を受け取り、キーワードが絞り込み条件として設定される。

ステップ104) 絞り込み条件設定部2において、絞り込み条件で指定されたキーワードのシソーラス辞書11上の位置を取得し、絞り込み条件設定部5においてその同義語、上位語、下位語を抽出し、照合対象語とし、照合対象語蓄積部10に登録する。

【0023】ステップ105) ハイパーリンク文書アクセス部4は、探索管理部5を介して、アクセス管理テーブル6に登録されたアクセス予定ノードへアクセスし、ハイパーリンク文書を取得して、取得文書蓄積部7に探索結果として保存すると共に、探索管理部5を介して、アクセス管理テーブル6の当該アクセス予定ノードの<処理結果>を『処理済』とする。

【0024】ステップ106) 探索管理部5は、取得した文書のノードの深さが最大ノード深さに達したかを調べ、最大ノード深さの場合には、ステップ109に移

行する。そうでない場合には、ステップ107に移行する。

ステップ107) 探索管理部5は、ハイパーリンク文書中から、リンク先に対応する説明テキストを抽出する。

【0025】ステップ108) アクセス候補絞り込み部3は、説明テキスト中に照合対象語が存在するかを調べ、存在する場合には、当該説明テキストに対応するリンク先をアクセス予定ノードとしてアクセス管理テーブル6へ登録する。

ステップ109) 探索管理部5は、アクセス管理テーブル中に『未処理』のアクセス予定ノードがあるか調べ、ある場合には、ステップ105に移行する。

【0026】ステップ110) ない場合には、探索結果表示部9において、探索結果出力要求を受け取ると、取得文書蓄積部7に蓄積された探索結果を出力する。

【0027】

【実施例】以下、図面と共に本発明の実施例を説明する。図5は、本発明の一実施例のハイパーリンク文書の例を示す。図5に示す、ハイパーリンク文書の[h0001]、[h0002]等は、ハイパーリンク文書に付与した識別子である。ここで示した文書(以下、ここでは、ハイパーリンク文書を省略して単に文書と記す)は、説明のため簡略化しており、表や画像等を含んだ文書であってもよい。ハイパーリンク文書の場合、表示書式等の指定を行うタグやリンク先を示すためのタグが定義されており、これらの情報を基にリンク先に対応する説明テキストを抽出できる。また、自然言語解析処理を行って、タグ情報のみでは抽出できない説明テキストに対しても抽出を行う構成としてもよい。

【0028】文書[h0001]では、文書[h0011]へのリンクに対し、「コンピュータ」が説明テキストに対応しており、文書[h0012]へのリンクに対し、「通販」が説明テキストに対応している。同様に、「企業」、「アート」、「メディア」は、各々[h0013]、[h0014]、[h0015]のリンクの説明テキストである。リンク先とそれに対応するテキストとを、

(<リンク先>、<説明テキスト>)

の形式で表すと、文書[h0001]には、リンク情報として、

(h0011, コンピュータ)

(h0012, 通販)

(h0013, 企業)

(h0014, アート)

(h0015, メディア)

が存在する。

【0029】図6は、本発明の一実施例のシソーラス辞書の例を示す。同図に示すシソーラス辞書11は、語が表す概念の関係を示す情報を蓄積した辞書であり、絞り

込みを行うための知識として使用される。同図の例では、両方向の矢印で同義語を表し、一方向の矢印で、上位・下位関係を表している。例えば、「販売」、「ショッピング」、「shopping」は、同義語であり、これらの語の下位語は、「通販」、「小売り」である。「小売り」から見ると、上位語は「販売」であり、下位語は、「百貨店」、「専門店」である。

【0030】ここでは、キーワードを指定して絞り込み条件とする一例として、シソーラス辞書11を参照して照合対象語に展開し、この照合対象語が説明テキスト中に含まれるか否かによって、リンク先を絞り込むよう構成した場合を例に説明する。本実施例では、アクセス候補絞り込み部3での参照が容易となるよう、絞り込み条件設定部2において、指定されたキーワードの同義語、上位語、下位語を照合対象語として照合対象蓄積部10へ保持する。

【0031】図7は、本発明の一実施例のアクセス管理テーブルの例を示す。同図(a)は、文書アクセス前のテーブルの状態を示し、(b)は、[h0001]文書取得後のテーブルの状態を示し、(c)は、[h0012]文書取得後のテーブルの状態を示す。以下、図4のフローチャートに従って、文書[h0001]が開始文書として設定され、最大ノード深さとして、“2”が設定された場合について説明する。

【0032】ステップ101) 処理要求受付部8より、探索条件設定の要求を受け取り、探索条件の初期条件として、開始文書が設定され、終了条件として、最大ノード深さが設定される。即ち、開始文書として文書[h0001]が設定され、最大ノード深さとして“2”が設定される。

ステップ102) 探索の初期条件として指定された開始文書をアクセス予定ノードとしてアクセス管理テーブル6に登録する。指定される開始文書は、複数個であってもよい。ここでは、図5の文書[h0001]が開始文書として指定された場合について説明する。開始文書[h0001]に対しては、図7(a)に示すように、＜アクセス予定ノード＞の項に「h0001」が、＜処理結果＞の項に「未処理」が、＜ノードの深さ＞の項に「0」が記述される。

【0033】ステップ103) 絞り込み条件設定部2において、処理要求受付部8からの、絞り込み条件設定要求を受け取り、キーワードが絞り込み条件として指定される。ここでは、キーワードとして、「ショッピング」と「家具」が指定されたものとする。

ステップ104) 絞り込み条件設定部2において、シソーラス辞書11を参照し、絞り込み条件で指定されたキーワードの同義語、上位語、下位語を抽出し、照合対象語とし、これを照合対象語蓄積部10へ保存する。図6のシソーラス辞書を使用した場合には、以下のように照合対象語が決定される。

【0034】「ショッピング」に対する照合対象語としては、同義語として、「販売」、「ショッピング」、「shopping」が抽出され、下位語として「通販」、「小売り」が抽出され、「小売り」の下位語として「百貨店」、「専門店」が抽出される。「家具」に対しては、上位語として、「商品」が、下位語として、「机」、「椅子」が抽出される。

【0035】即ち、照合対象語は、「販売」、「ショッピング」、「shopping」、「通販」、「小売り」、「百貨店」、「専門店」、「商品」、「机」、「椅子」となり、これらの語が照合対象語蓄積部10に蓄積される。ステップ105) アクセス管理テーブル6に登録されたアクセス予定ノードにおいて、＜処理結果＞が『未処理』のノードへネットワーク12を介してアクセスし、ハイパーリンク文書を取得する。そして、探索結果として取得文書蓄積部7へ取得したハイパーリンク文書を保存すると共に、当該アクセス予定ノードの＜処理結果＞を『処理済』とする。ここでは、説明の簡略化のため、リンク先をハイパーリンク文書の識別子のみで表しているが、リンク先としてハイパーリンク文書の所在を表す情報、即ち、どのサーバに存在するかという情報も含んでおり、どのサーバへアクセスすればよいかは決定できる。

【0036】ステップ106) 取得した文書のノード深さが最大ノード深さに達したかを調べる。アクセス管理テーブル6の＜ノード深さ＞を検査することにより、最大ノード深さに達したかどうかを調べることができる。最大ノード深さに達した場合には、ステップ107、ステップ108をスキップして、ステップ109に移行する。

【0037】ステップ107) 取得したハイパーリンク文書中から、リンク先に対応する説明テキストを抽出する。文書[h0001]の場合には、(h0011, コンピュータ)、(h0012, 通販)、(h0013, 企業)、(h0014, アート)、(h0015, メディア)が抽出される。

【0038】ステップ108) 説明テキスト中に照合対象語が存在するかを調べ、存在する場合には、当該説明テキストに対応するリンク先をアクセス予定ノードとしてアクセス管理テーブル6へ登録する。文書[h0001]の場合には、「通販」のみが、(h0012, 通販)として存在する。そこで、文書[h0012]をアクセス管理テーブル6へ登録する。このとき、＜処理結果＞に『未処理』を記述し、＜ノードの深さ＞は親のノードの深さに1を加算したものを記述する。文書[h0001]の＜ノードの深さ＞は、0であるので、文書[h0012]の＜ノードの深さ＞は1となる。その結果、アクセス管理テーブル6の内容は、図7(b)に示す通りである。

【0039】ステップ109) アクセス管理テーブル

6中に、『未処理』のアクセス予定ノードがあるかを調べ、ある場合にはステップ105に移行し、『未処理』のアクセス予定ノードがない場合には、ステップ110へ移行する。ステップ110) 処理要求受付部8より、探索結果出力要求を受け取ると、探索結果表示部9は、取得文書蓄積部7に蓄積された探索結果を処理要求受付部8に出力する。

【0040】上記で説明した処理により、図5に示すハイパーリンク文書の場合には、ステップ105からステップ110に至るまでの処理は、以下のようになる。

(1) 初期条件の開始文書として指定された文書[h0001]が取得される(ステップ105)。

(2) 文書[h0001]の<ノード深さ>は、0であり、最大ノード深さ2に達していないため、次のステップ107に移行する(ステップ106)。

【0041】(3) 文書[h0001]から、説明テキスト(h0011、コンピュータ)、(h0012、通販)、(h0013、企業)、(h0014、アート)、(h0015、メディア)が抽出される(ステップ107)。

【0042】(4) 照合対象語の「通販」が(h0012、通販)に存在するので、[h0012]をアクセス予定ノードとしてアクセス管理テーブル6に登録する。

(5) [h0012]が未処理なので、ステップ105に移行する(ステップ109)。

(6) 文書[h0012]が取得される(ステップ105)。

【0043】(7) 文書[h0012]の<ノード深さ>は1であり、最大ノード深さ2に達していないため、次のステップ107に移行する(ステップ106)。

(8) 文書[h0012]から、説明テキスト(h0121、百貨店)、(h0122、ショッピングモール)、(h0123、食品)、(h0124、コンピュータ)、(h0125、書籍)が抽出される(ステップ107)。

【0044】(9) 照合対象語「百貨店」が(h0121、百貨店)に存在し、照合対象語「ショッピング」は(h0122、ショッピングモール)の文字列“ショッピングモール”中に含まれているので、[h0121]及び[h0122]がアクセス予定ノードとしてアクセス管理テーブル6に登録される。ここで、[h0121]及び[h0122]が未処理なので、ステップ105へ移行する(ステップ109)。

【0045】(10) [h0121]及び[h0122]が未処理なので、ステップ105に移行する(ステップ109)。

(11) 文書[h0121]が取得される(ステップ105)。

(12) 文書[h0121]の<ノード深さ>は“2”であり、最大ノード深さ“2”に達しているため、ステップ107、ステップ108をスキップして、ステップ109に移行する(ステップ106)。

【0046】(13) [h0122]が『未処理』として残っているので、ステップ105に移行する(ステップ109)。

(14) 文書[h0122]が取得される(ステップ105)。

10 (15) 文書[h0122]の<ノード深さ>は、“2”であり、最大ノード深さ“2”に達しているため、ステップ107、ステップ108をスキップして、ステップ109に処理が移行する(ステップ106)。

【0047】(16) 未処理として残っているアクセス予定ノードは存在しないので、ステップ110に移行する(ステップ109)。

(17) 探索結果出力の要求があると、取得文書蓄積部7に蓄積された文書[h0001]、[h0012]、[h0121]、[h0122]が出力される。

20 【0048】このようにして、絞り込まれたリンク先のみへアクセスが行われる。上記の例のように、例えば、「家具を買いだめしたい」という動機によって、「ショッピング」及び「家具」を指定すれば、ユーザが所望する情報が含まれる可能性のある文書以外へのアクセスが抑制される。このため、最大ノード深さを大きくとっても合理的な時間で、目的にあった情報を効率的に収集できる。

【0049】なお、本発明は、上記の実施例に限定されことなく、特許請求の範囲内で種々変更・応用が可能である。

30 【0050】

【発明の効果】上述のように、本発明の情報収集方法及び装置によれば、絞り込み条件でリンク先を絞り込むため、アクセスする文書量を大幅に削減でき、所望する情報が含まれる可能性が高い文書を容易に取得することができる。このため、本発明による装置を用いることにより、効率的に情報収集を行うことができる。

【0051】従来の装置との比較を定量的に行うと以下のようになる。従来の装置では、すべてのリンク先に対してアクセスを行い、文書を取得する。このため、1文書あたりn個のリンク先があり、m階層までアクセスを行うとすると、m階層目の文書個数は、nのm乗個となり、探索の深さを大きくすると、膨大な量の文書をアクセスすることになる。

【0052】これに対し、本発明の情報収集装置では、リンク先絞り込み条件によりリンク先を絞り込むため、アクセスする文書数を削減できる。具体的には、1文書当たり、1/kに絞り込まれたとすると、m階層目にアクセスする文書数は、1/(kのm乗)に削減される。例えば、1/3に絞り込まれたとすると、5階層目では、絞り込みを行わない場合に比べ、1/243に削減

される。

【図面の簡単な説明】

【図1】本発明の原理を説明するための図である。

【図2】本発明の原理構成図である。

【図3】本発明の情報収集装置の構成図である。

【図4】本発明の情報収集装置の動作のフローチャートである。

【図5】本発明の一実施例のハイパーリンク文書の例である。

【図6】本発明の一実施例のシソーラス辞書の例である。

【図7】本発明のアクセス管理テーブルの例である。

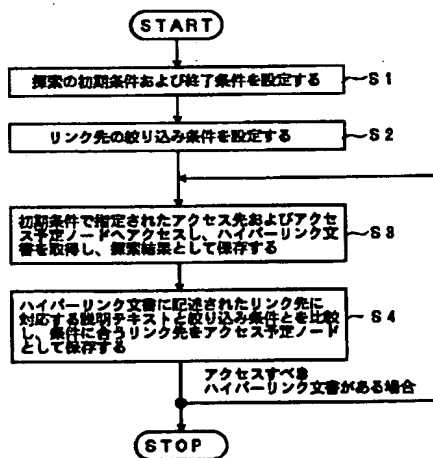
【図8】従来の情報収集装置の構成図である。

【図9】従来装置の処理のフローチャートである。

【符号の説明】

【図1】

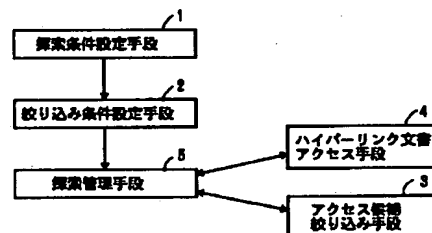
本発明の原理を説明するための図



- 1 探索条件設定手段、探索条件設定部
- 2 絞り込み条件設定手段、絞り込み条件設定部
- 3 アクセス候補絞り込み手段、アクセス候補絞り込み部
- 4 ハイパーリンク文書アクセス手段、ハイパーリンク文書アクセス部
- 5 探索管理手段、探索管理部
- 6 アクセス管理テーブル
- 7 取得文書蓄積部
- 8 処理要求受付部
- 9 探索結果表示部
- 10 照合対象語蓄積部
- 11 シソーラス辞書
- 12 ネットワーク
- 13 サーバ群

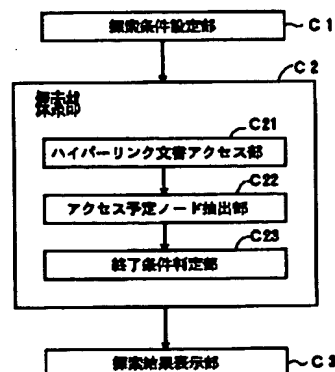
【図2】

本発明の原理構成図

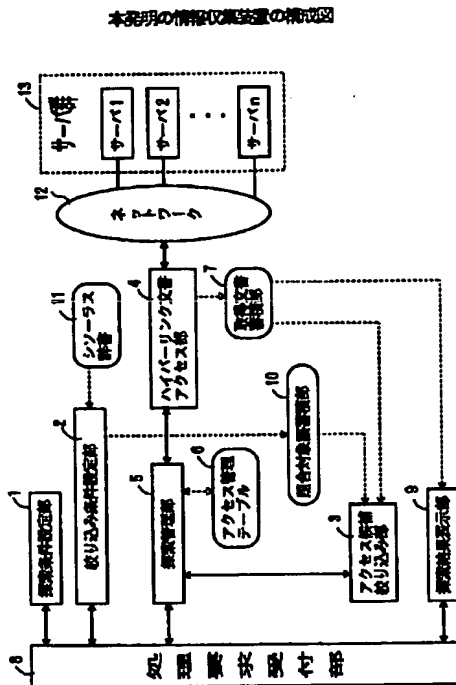


【図8】

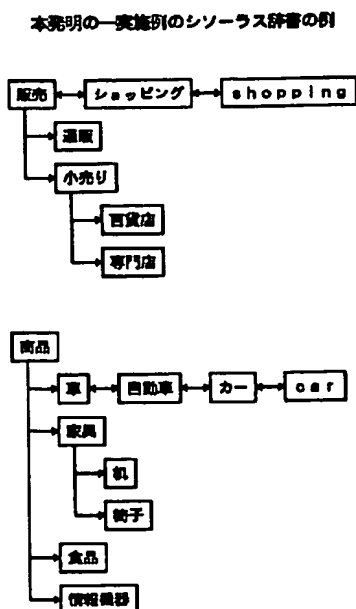
従来の情報収集装置の構成図



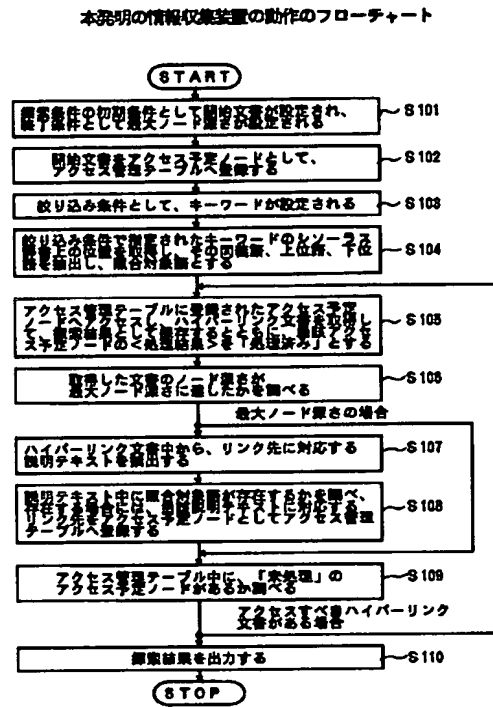
【図 3】



【図6】

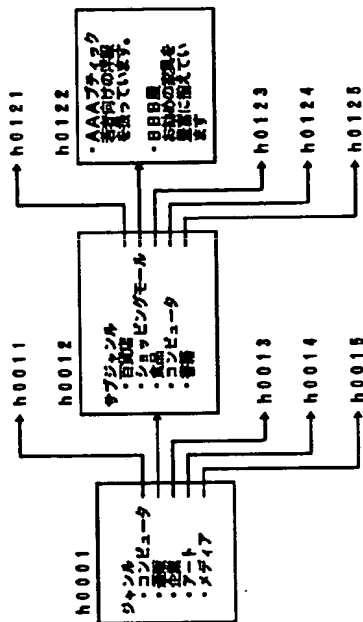


【図4】



【図5】

本発明の一実施例のハイパーリンク文書の例



【図7】

本発明の一実施例のアクセス管理テーブルの例

アクセス予定 ノード	処理結果	ノードの深さ
h0001	未処理	0

(a) 文書アクセス前

アクセス予定 ノード	処理結果	ノードの深さ
h0001	処理済み	0
h0012	未処理	1

(b) h0001文書取得後

アクセス予定 ノード	処理結果	ノードの深さ
h0001	処理済み	0
h0012	処理済み	1
h0121	未処理	2
h0122	未処理	2

(c) h0012文書取得後

【図9】

従来装置の処理のフローチャート

